Resource-rational reinforcement learning and sensorimotor causal states

Sarah Marzen
W. M. Keck Science Department
Pitzer, Scripps, and Claremont McKenna College
(Dated: April 30, 2024)

We propose a new computational-level objective function for theoretical biology and theoretical neuroscience that combines: reinforcement learning, the study of learning with feedback via rewards; rate-distortion theory, a branch of information theory that deals with compressing signals to retain relevant information; and computational mechanics, the study of minimal sufficient statistics of prediction also known as causal states. We highlight why this proposal is likely only an approximation, but is likely to be an interesting one, and propose a new algorithm for evaluating it to obtain the newly-coined "reward-rate manifold". The performance of real and artificial agents in partially observable environments can be newly benchmarked using these reward-rate manifolds. Finally, we describe experiments that can probe whether or not biological organisms are resource-rational reinforcement learners.

I. INTRODUCTION

According to Marr, understanding biological organisms entails uncovering three levels: the computational, the algorithmic, and the mechanistic [1, 2]. At the computational level, we ask what organisms are trying to do. What objective function might they be using? At the algorithmic level, we ask what algorithm they are using to accomplish that objective. And at the mechanistic level, we ask how they are implementing that algorithm in their messy hardware. None of these levels have been completely understood in theoretical neuroscience or theoretical biology, despite major advances such as the Hodgkin-Huxley model that describes how neurons behave using electrical engineering ideas.

In this manuscript, we claim that resource-rational decision making is a plausible first attempt at the computational level [3]. This research program goes by the name of computational rationality [4], rational inattention [5, 6], and many other names. The basic idea behind it is that organisms endeavour to solve tasks as well as possible, but are limited in their ability to solve tasks by various resources. These resources can be time limitations, memory limitations, material limitations, or other limitations.

There is much debate over how to implement resource-rational decision making quantitatively, but information-theoretic codings of resources [7] and reinforcement learning-based measures of the quality of decision making [8] might be the key to understanding the full sensorimotor loop. Already, reinforcement learning has been famously used to describe dopaminergic signals [9], although there is much recent debate over whether or not that mechanistic level description is appropriate [10]. On the other hand, using information-theoretic quantities as perceptual costs has allowed researchers to explain a number of empirical findings in a wide variety of areas in the last two decades, including various aspects of macroeconomic behavior [5, 6], Shepard's universal law of generalization [11], the fuzziness of color naming systems [12], sub-optimal prediction in sequence learning [13], and a number of empirical findings on neural coding and working memory [14]. And, while not done on humans, recent work has shown that salamander retinal ganglion cells [15] and cultured cortical neurons from rats [16] both predict stimuli efficiently in an information-theoretic sense but do not always predict well in an absolute sense. Information-theoretic costs can be justified both using material constraints [7] and nonequilibrium thermodynamics [17, 18].

There have been attempts to combine information theoretic resource constraints and reinforcement learning objectives in Refs. [19–21], but in this manuscript, we will argue that these attempts require combination to achieve the correct objective. We will give a new Blahut-Arimoto-like algorithm for calculating what we call the "reward-rate manifold", which describes how well an organism (real or artificial) can attain reward under the information-theoretic resource constraints. In order to provide an algorithm, we will prove that the sensorimotor causal states of Ref. [19] can replace semi-infinite histories of observations and actions, essentially making it possible to calculate an infinite object with finite resources.

We begin by describing the new proposed objective function, continue by providing an algorithm to efficiently calculate the newly-described reward-rate manifold, and finish by showing an example reward-rate manifold. We conclude by describing what might be done with this contribution.

II. A NEW COMPUTATIONAL-LEVEL OBJECTIVE FOR THEORETICAL BIOLOGY

We start by discussing proposals for a computational-level objective for theoretical biology in Sec. II A and move to introducing my own in Sec. II B. The environment under consideration is known in reinforcement learning [8] as a Partially Observable Markov Decision Process (POMDP), in which there is an underlying Markov state w describing the environment, actions

a that describe what the agent can do, noisy and partial observations o of the underlying world state w that describe what the agent sees, a discount factor γ that describes how agents treat future rewards, and a reward function $r(w, a)$ that describes how much "reward" an agent receives when the world is in state w and the agent takes action a. These rewards can take the form of food, shelter, sleep, and so on, and are left unspecified for the purpose of this paper. In an experiment, one might imagine giving rats sugar or humans money.

A. Attempts So Far

The first instance of such an objective function incorporating sensors and actuators is perhaps a paper by Still [19]. She imagined that an organism sees observations o_t at time t, converts past actions and observations to sensory state s_t, and takes action a_t right after based also on that history. The history of observations and actions is labeled h_t and the future of observations is labeled z_t. She imagines that h_t is used to inform both s_t and a_t separately. Still suggests that one should try to maximize $I[s, a; z] - \lambda I[s; h] - \mu I[a; h]$ where $\lambda,\ \mu$ are Lagrange multipliers and time indices have been dropped for easier-to-read notation. In Ref. [19], Still found optimal sensors to be sensorimotor causal states (described in Sec. III) in the limit that $\lambda \to 0$ and also identified optimal action policies in the limit that $\mu \to 0$.

The first term in this objective is interesting, but maximizing this term usually leads to large periodic loops when $\lambda,\ \mu$ are near enough to 0. (Large periodic loops have a high mutual information between past and future.) That is unfortunately a limit of interest for higher-level organisms that can pick up the aforementioned sensorimotor causal states. Although some work [22] claims that these high predictive information processes correspond to processes that learn underlying parameters of the environment model, that is only true in a nonergodic case [23]. It may be possible in certain environments to see something more complex [24]. For lower-level organisms, the limit $\lambda,\ \mu \to \infty$ is of greater interest, but that leads to sensory states and actions that depend not at all on the history and are instead biased coin flips, by simulations not shown here. A quick theoretical argument suggests that should be the case– $I[s; h],\ I[a; h]$ can both be set to 0 if $s,\ a$ have no dependence on h.

The next instance of such an instantiation that is information-theoretic comes identically from Ref. [20] and Ref. [21]. Here, the information-theoretic term $I[s, a; z]$ is replaced by the usual reinforcement learning term V_π, the sum total of discounted rewards. Rewards depend on the underlying Markov state of the environment w_t, so that $V_\pi = \sum_t \gamma^t r(w_t, a_t)$ where γ is a discount factor, r the usual reward function [8], and w_t and a_t the world state and actions at time t. It is straightforward to generalize to continuous-time by introducing an integral. There is no cost for complicated sensory states s, unlike in Ref. [19]. There is only a cost on transmitting information from sensory state to actions $I[s; a]$. As a result, the objective function reads $V_\pi - \beta I[s; a]$. Note that here, s is used to inform the action a rather than the entire history h being used to inform the action. This rings more true to neuroscience, as we describe in the next section.

Finally, from Ref. [25], it is clear that bacteria are reinforcement learners that choose a strategy that works best for the worst-case scenario rather than operating on a discounted sum of rewards. We ignore these lower-level organisms for the purpose of this paper, conjecturing that you see such minimax behavior only when the organism lacks a theory of mind.

B. The New Objective

In a way, what we propose is an instantiation of the ideas in Ref. [26] but for POMDPs. We must carefully decide which terms to include in the final objective function describing an organism trying to navigate a sensorimotor feedback loop. Altogether, we would like an objective function that naturally balances exploration and exploitation, meaning that an organism should explore its environment naturally before exploiting the information it has obtained to survive; and we would like an objective function that includes as many resource constraints as possible. A simple combination of the objective functions that exist so far as mentioned in Sec. II A yields:

$$\mathcal{L} = V_\pi - \beta I[s; a] - \lambda I[h; s] \tag{1}$$

where $\beta,\ \lambda$ are constants. This is really the unconstrained version of a constrained objective function:

$$R(MI_{s,a}, MI_{h,s}) = \max_{I[s;a] \leq MI_{s,a}, I[h;s] \leq MI_{h,s}} V_\pi \tag{2}$$

so that $\beta,\ \lambda$ are Lagrange multipliers and $MI_{s,a}$ and $MI_{h,s}$ are adjustable constants.

With the constrained objective function, we define the reward-rate manifold, in which $MI_{h,s}$ is on the x-axis, $MI_{s,a}$ is on the y-axis, and V_π on the z-axis. The manifold separates achievable combinations of information-theoretic rates $I[h; s],\ I[s; a]$ and rewards V_π and unachievable combinations, as in rate-distortion theory [7] and predictive rate-distortion theory [27]. In other words, the reward-rate manifold defines a Pareto front. We show an example of this reward-rate manifold in Sec. **??**.

First, we discuss the term that allows the organism to accumulate reward. The term V_π naturally implies that we must both explore and exploit: to reap rewards, one must survey all available options (within reason) and choose the best one rather than merely sticking with the first good option that comes around. However, much effort has been spent in reinforcement learning trying to add additional terms or alter action policies so that a better balance of exploration and exploitation is achieved, e.g. as in Ref. [28].

Next, we discuss the information-theoretic resource term that suggests the organism should aim for a simpler actuator. We must convey the sensory state s to find the action policy a using the conditional probability $\pi(a|s)$ that signifies the action policy [8]– the actuator a does not have direct access to histories h– and so $I[s;a]$ is the appropriate term, as identified by Refs. [20, 21].

Finally, we discuss the information-theoretic resource term that suggests the organism should aim for a simpler sensory layer [15]. If we think about the human brain, observations from the retina o must combine with efference copies a at V1 to give us a sensory state s that can be used to determine actions. Mathematically, there is some input-dependent dynamical system that takes in information from the efference copy and the observations and turns it into something that is not quite the history h written down by Still, but has information going back to the beginning of when the organism has opened its eyes. Hence we are perhaps somewhat justified in replacing this variable by h. This information must be communicated to the next layer in the brain, justifying $I[h;s]$ as the next resource constraint.

Biology is not likely to directly work on this objective function, but might be subject to resource constraints that force it to essentially maximize this objective function. Essentially, the resource constraints that biology operates on might look more like material constraints [29] or energy constraints [17, 18], both which lead to mutual informations as the natural stand-in using results from information theory or nonequilibrium thermodynamics. See App. A.

III. MAKING IT CALCULABLE WITH SENSORIMOTOR CAUSAL STATES

Sensorimotor causal states as defined in Ref. [19] are usually also belief states of the POMDP [30]. Belief states are the probability distribution over the underlying Markov state of the environment (or more technically, of the POMDP) w given the history h, and one uses these to "solve" the POMDP– to determine one's action policy [30, 31].

These sensorimotor causal states come from a coarse-graining relationship, as in Ref. [19, 32]. Take histories h and consider two histories h, h' equivalent if $P(w|h) = P(w|h')$. Note the difference from Ref. [19]– we have replaced future observations with the underlying Markov state of the POMDP. The best guide to the future of the observations is the underlying Markov state of the environment w. This is unobtainable directly, so in any real algorithm to ascertain sensorimotor causal states, one might use the future of observations instead. Regardless, the clusters of histories are labeled σ, sensorimotor causal states, and the sensorimotor causal state to which history h belongs is given by $\epsilon^+(h)$. We define sensorimotor causal states in this modified way so that the proof of the main theorem in this paper is clear; as an added benefit, these modified sensorimotor causal states are now *exactly* the belief states.

The objective function from the previous section was $V_\pi - \beta I[s;a] - \lambda I[h;s]$. We can replace histories h with sensorimotor causal states σ if we wish to find statistics of good sensors [7] or to calculate the reward-rate manifold. (Importantly, the obtained sensor $p(s|h)$ and actuator $\pi(a|s)$ from maximizing this objective might not be good sensors or actuators themselves by the original material constraints [7].) To prove this, note that there is no change to V_π or $I[s;a]$ if sensory states $p(s|h)$ are recoded as $p(s|\sigma = \epsilon^+(h))$, similar to what is true in Ref. [27]. And, as in Ref. [27], $I[s;h] = I[s;\sigma] + I[s;h|\sigma]$ only decreases with this recoding to $I[s;\sigma]$ since $I[s;h|\sigma] \geq 0$. The objective function therefore benefits from this recoding. As a result, as expected, it is optimal to pick up sensorimotor causal states using the recurrent neural network that governs the sensory layer in biology.

The new insight into sensory states is that they should pick up nothing else, however lossy; and that the objective function can be rewritten with histories h replaced with sensorimotor causal states σ.

We now specialize to the case of no discount factor $\gamma = 1$, in which case V_π turns into a sum of rewards. For a POMDP, one can define a reward function on belief states σ and actions a from the underlying reward function on underlying Markov states of the environment w and actions a [30], but we avoid this step. (It is not necessary for calculating the reward-rate manifold for the experiments we plan to do in the future.) Under a stationarity condition, V_π turns into $T\langle r(w,a)\rangle_{p(w,a)}$, where T is the total number of time steps in the organism's life. We can ignore the additional factor of T by rescaling β, λ.

In this case, from Appendix B, we can calculate the reward-rate manifold by using the iterative algorithm which updates $\pi_n(a|s)$ and $p_n(s|\sigma)$ as in the usual information bottleneck algorithm [33]:

$$\pi_{n+1}(a|s) \;=\; \pi_n(a)\frac{\exp\left(\frac{1}{\beta}\sum_{\sigma,w} p_n(\sigma|s)p(w|\sigma)r(w,a)\right)}{Z_{\beta,n}(s)} \tag{3}$$

where $Z_{\beta,n}(a)$ is a partition function or normalization factor, similar to Refs. [19, 33], so that

$$Z_{\beta,n}(a) \;=\; \sum_a \pi_n(a)\exp\left(\frac{1}{\beta}\sum_{\sigma,w} p_n(\sigma|s)p(w|\sigma)r(w,a)\right). \tag{4}$$

Similar manipulations for $p(s|\sigma)$ gives

$$p_{n+1}(s|\sigma) = \frac{p_n(s) \exp\left(\frac{1}{\lambda} \sum_{a,w} \pi_n(a|s) p(w|\sigma) r(w,a)\right)}{Z_{\lambda,n}(\sigma)} \tag{5}$$

where $Z_{\lambda,n}(\sigma)$ is again a partition function or normalization factor,

$$Z_{\lambda,n}(\sigma) \;=\; \sum_s p_n(s) \exp\left(\frac{1}{\lambda} \sum_{a,w} \pi_n(a|s) p(w|\sigma) r(w,a)\right). \tag{6}$$

When environments and the number of possible actions are large, it is profitable to avoid numerical errors in coding by taking the logarithm of both sides of Eqs. 3 and 5.

As λ, β change from 0 to ∞, we trace out the entire two-dimensional reward-rate manifold. Because the objective function is convex in the sensor description $p(s|\sigma)$ and actuator description $\pi(a|s)$, this generalized Blahut-Arimoto algorithm will converge to the global optimum as $n \to \infty$. As in Ref. [19], in the limit that $\lambda \to 0$, we find that s recovers exactly the sensorimotor causal states σ and in the limit that $\beta \to 0$, we find a deterministic action policy. However, the goal here is not necessarily to find sensors or actuators– though by conjecture *statistics* of good ones can be obtained from this algorithm [7]– but to calculate a reward-rate manifold so as to benchmark how well biological and artificial agents reap reward under resource constraints in POMDP environments.

Note that before this theorem, operating on long histories to calculate the reward-rate manifold would encounter a curse of dimensionality based on the length of the history. We have replaced histories with sensorimotor causal states, bypassing this curse of dimensionality [34], as in Ref. [27]. If the POMDP is somehow known, it is possible to calculate the reward-rate manifold. Code is available on GitHub, and an example manifold is shown in Fig. 1.

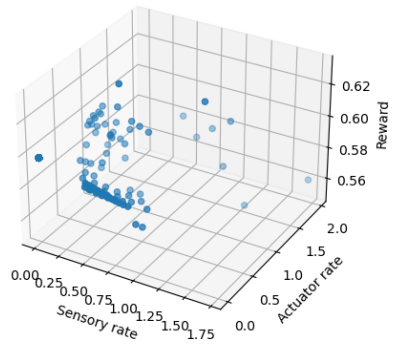

FIG. 1. An example of a reward-rate function obtained using the generalized Blahut-Arimoto algorithm when the number of possible actions is 50, the number of sensorimotor causal states is 40, the number of underlying environmental states is 20, the reward function is chosen uniformly at random from the unit interval, the probability distribution over causal states is chosen at random from a Dirichlet distribution with concentration parameter $\alpha = 1$, and the conditional probability distributions over underlying environmental states given causal states are chosen at random from Dirichlet distributions with concentration parameter $\alpha = 1$. Sweeps were done over β, λ between 0.01 and 1 with 100 points on each grid. One errant point is observed at a low sensory rate $MI_{h,s}$ and low actuator rate $MI_{s,a}$ perhaps due to lack of convergence of the algorithm. Only 8000 iterations were used per point. A phase transition is otherwise observed from no rate and some reward to nonzero rate and more reward as expected from Ref. [27].

IV. CONCLUSION

In this manuscript, we have proposed a new computational-level objective function for theoretical biology and theoretical neuroscience that combines: reinforcement learning [8], the study of learning with feedback via rewards; rate-distortion theory, a branch of information theory [7, 33] that deals with compressing signals to retain relevant information; and computational mechanics, the study of minimal sufficient statistics of prediction also known as causal states [19, 32]. We have highlighted why this proposal is likely only an approximation, but is likely to be an interesting one, and proposed a new algorithm for evaluating it to obtain the newly-coined "reward-rate manifold".

It is important to stress that biological organisms are likely not operating directly on this objective function. Rather, they are naturally subject to resource constraints that lead to them naturally maximizing this objective function. Nor are the sensors and actuators revealed by this objective function likely to be the actual sensors and actuators used– famously, the sensors and actuators that are revealed only provide statistics that describe the true sensors and actuators that do well on the objective function [7].

In order to calculate this reward-rate manifold, it will usually be necessary to use the sensorimotor causal states

first proposed in Ref. [19]. One might reasonably ask why the organism should have access to the sensorimotor causal states. Rather, the organism is likely trying to infer sensorimotor causal states using some algorithm that we have not yet determined [31, 34]. As in Refs. [13, 15, 16], we envision a raft of experiments that involve the experimentalist knowing the environmental statistics with which the organisms are probed and using their knowledge of sensorimotor causal states to calculate the reward-rate manifold, calculate the reward and rates of the organism from behavioral and neural data, and then place the organism's operation relative to the reward-rate manifold as is common in rate-distortion theory [7]. This will enable a stringent test of whether or not the organism really is maximizing expected reward subject to information-theoretic rate constraints.

This is likely only a first approximation to the true computational level objective. Future efforts might focus on including time, as much effort has been spent understanding the speed-accuracy-energy tradeoff in nonequilibrium thermodynamics [35]. Also, this proposal does not solve at all the algorithmic or mechanistic level. Those are left to methods such as maximum likelihood determination of the true sensory and actuator strategies [36, 37]. Still, we hope that this contribution allows for the development of a research program that will finally unfurl the computational level of theoretical biology and theoretical neuroscience.

ACKNOWLEDGMENTS

I would like to thank Dmitri Chklovskii and Rainer Engelken for inspiring conversations.

Appendix A: Reasoning for Mutual Informations From the Rate-Distortion Theorem

Before I describe the resource constraints for this POMDP, let us describe the rate-distortion theorem [7]. It will justify why material constraints can be replaced by mutual informations.

In the classic rate-distortion setup, one sends a sequence of n letters $x_{0:n}$ to an encoder that chooses one of M words for those n letters and then sends that word to a decoder which produces a guess as to what those letters were, $\hat{x}_{0:n}$. The material constraint is actually $\log M/n$, not a mutual information. This corresponds to a more intuitive notion of resource constraints in the biological sense– number of molecules or number of neurons, normalized by "blocklength" n. Some distortion measure is defined, $d(x, \hat{x})$, which in some extensions can be a distortion of the entire block $x_{0:n}$ relative to $\hat{x}_{0:n}$ rather than letter-by-letter. There are some rates $\log M/n$ and distortions $\sum d(x_i, \hat{x}_i)/n$ that are achievable and some that are unachievable given any combination of encoder and decoder. A theorem shows that the curve separating achievable from unachievable is given by replacing the rate $\log M/n$ with a mutual information $I[X; \hat{X}]$ and the average distortion with an expected distortion if all is memoryless. This curve is accurate in the limit that blocklength n goes to infinity. Otherwise, the rate-distortion curve that separates achievable from unachievable is given by $R_n(D)$ rather than $R(D)$, and $R_n(D)$ is horribly difficult to calculate [7]. In essence, what I will try to argue is that biology is in the limit of very large n sometimes, and so it is okay to use mutual informations to calculate the "reward-rate manifold"– the two-dimensional manifold that separates allowable from unallowable combinations of the two rates to be discussed and the reward V_π. Otherwise, $R_n(D)$ places an upper bound on $R(D)$, and since the reward is the flip of the distortion, the corresponding logic is that $R_n(MI_{s,a}, MI_{h,s})$ places a lower bound on $R(MI_{s,a}, MI_{h,s})$.

The key material constraint that we wish to think about is the number of neurons, either in the sensory layer or in the actuator layer. If there is a combinatorial code, then the number of words M is equivalent to 2^{num} where num is the number of neurons. A resource constraint that is reasonable is therefore $\log M$. This must be modulated by a blocklength– some sense of timescales. The NMJ (neuromuscular junction, or actuator layer) is thought to operate by a rate code, while the sensory layers are thought to operate on sub-millisecond timescales [38] and the environment is thought to operate on extremely large timescales given that naturalistic video is described by power laws. Given all this, the effective blocklength for the actuators is likely to be very high, so that $I[s; a]$ is justified; and $I[h; s]$ provides us with a lower bound on the reward-rate function.

A complication exists: the environment is memoryful, and so are the sensors and actuators. Typically memoryful processes have incalculable objectives [7]. As a result, I am replacing material constraints with mutual informations by conjecture as an approximation to what is likely true.

Finally, Landauer bounds suggest that mutual informations might replace work [39].

Appendix B: Derivation of a Generalized Blahut-Arimoto Algorithm

I start with the unconstrained objective function

$$\mathcal{L} = \langle r(w_t, a_t) \rangle - \beta I[s_t; a_t] - \lambda I[\sigma_t; s_t] - \gamma_s \sum p(\sigma_t) p(s_t|\sigma_t) - \gamma_a \sum p(s_t) p(a_t|s_t) \tag{B1}$$

for discrete state spaces. I take partial derivatives with respect to $p(a_t|s_t)$ and set them equal to 0. First:

$$\begin{aligned}
\frac{\partial r(w_t, a_t)\rangle}{\partial p(a_t|s_t)} &= \frac{\partial}{\partial p(a_t|s_t)} \sum p(w_t, a_t) r(w_t, a_t) && \text{(B2)} \\
&= \frac{\partial}{\partial p(a_t|s_t)} \sum p(w_t, a_t, s_t, \sigma_t) r(w_t, a_t) && \text{(B3)} \\
&= \frac{\partial}{\partial p(a_t|s_t)} \sum p(a_t|s_t) p(s_t|\sigma_t) p(w_t|\sigma_t) p(\sigma_t) r(w_t, a_t) && \text{(B4)} \\
&= \sum p(\sigma_t) p(s_t|\sigma_t) p(w_t|\sigma_t) r(w_t, a_t). && \text{(B5)}
\end{aligned}$$

Second:

$$\frac{\partial I[s_t; a_t]}{\partial p(a_t|s_t)} = \frac{\partial}{\partial p(a_t|s_t)} \left(H[a_t] - H[a_t|s_t] \right) \tag{B6}$$

where

$$\begin{aligned}
\frac{\partial H[a_t|s_t]}{\partial p(a_t|s_t)} &= -\frac{\partial}{\partial p(a_t|s_t)} \sum p(s_t) p(a_t|s_t) \log p(a_t|s_t) && \text{(B7)} \\
&= -p(s_t) \left(1 + \log p(a_t|s_t) \right) && \text{(B8)}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial H[a_t]}{\partial p(a_t|s_t)} &= -\frac{\partial}{\partial p(a_t|s_t)} \sum p(a_t) \log p(a_t) && \text{(B9)} \\
&= -\sum (1 + \log p(a)) \frac{\partial p(a)}{\partial p(a_t|s_t)} && \text{(B10)} \\
&= -\sum \delta_{a,a_t} p(s_t) (1 + \log p(a)) && \text{(B11)} \\
&= -p(s_t) \left(1 + \log p(a_t) \right) && \text{(B12)}
\end{aligned}$$

which means

$$\begin{aligned}
\frac{\partial I[s_t; a_t]}{\partial p(a_t|s_t)} &= -p(s_t) \left(1 + \log p(a_t) \right) + p(s_t) \left(1 + \log p(a_t|s_t) \right) && \text{(B13)} \\
&= p(s_t) \log \frac{p(a_t|s_t)}{p(a_t)}. && \text{(B14)}
\end{aligned}$$

Third:

$$\frac{\partial I[s_t; \sigma_t]}{\partial p(a_t|s_t)} = 0. \tag{B15}$$

Fourth:

$$\frac{\partial \sum p(a_t|s_t)}{\partial p(a_t|s_t)} = 1 \tag{B16}$$

and finally the last partial derivative is 0. This gives

$$0 = \sum_{\sigma_t, w_t} p(\sigma_t)p(s_t|\sigma_t)p(w_t|\sigma_t)r(w_t, a_t) - \beta p(s_t) \log \frac{p(a_t|s_t)}{p(a_t)} - \gamma_a p(s_t) \tag{B17}$$

$$\beta p(s_t) \log \frac{p(a_t|s_t)}{p(a_t)} = \sum_{\sigma_t, w_t} p(\sigma_t)p(s_t|\sigma_t)p(w_t|\sigma_t)r(w_t, a_t) - \gamma_a p(s_t) \tag{B18}$$

$$\log \frac{p(a_t|s_t)}{p(a_t)} = \frac{1}{\beta p(s_t)} \sum_{\sigma_t, w_t} p(\sigma_t)p(s_t|\sigma_t)p(w_t|\sigma_t)r(w_t, a_t) - \frac{\gamma_a}{\beta} \tag{B19}$$

$$\frac{p(a_t|s_t)}{p(a_t)} = \exp\left(\frac{1}{p(s_t)} \sum_{\sigma_t, w_t} p(\sigma_t)p(s_t|\sigma_t)p(w_t|\sigma_t)r(w_t, a_t) - \frac{\gamma_a}{\beta}\right) \tag{B20}$$

$$= \exp\left(\frac{1}{\beta} \sum_{\sigma_t, w_t} p(\sigma_t|s_t)p(w_t|\sigma_t)r(w_t, a_t) - \frac{\gamma_a}{\beta}\right) \tag{B21}$$

$$p(a_t|s_t) = p(a_t) \frac{\exp\left(\frac{1}{\beta} \sum_{\sigma_t, w_t} p(\sigma_t|s_t)p(w_t|\sigma_t)r(w_t, a_t)\right)}{Z_\beta(s_t)} \tag{B22}$$

where $Z_\beta(a_t)$ is the partition function or normalization factor. Similar manipulations for $p(s_t|\sigma_t)$ gives

$$p(s_t|\sigma_t) = \frac{p(s_t) \exp\left(\frac{1}{\lambda} \sum_{a_t, w_t} p(a_t|s_t)p(w_t|\sigma_t)r(w_t, a_t)\right)}{Z_\lambda(\sigma_t)} \tag{B23}$$

where $Z_\lambda(\sigma_t)$ is the partition function or normalization factor. To retrieve the generalized Blahut-Arimoto algorithm for the two-dimensional rate-reward manifold, we simply take Eqs. B22 and B23 and iterate them.

[1] D. Marr, Vision new york: Freeman, (1982).

[2] D. Levenstein, V. A. Alvarez, A. Amarasingham, H. Azab, Z. S. Chen, R. C. Gerkin, A. Hasenstaub, R. Iyer, R. B. Jolivet, S. Marzen, *et al.*, On the role of theory and modeling in neuroscience, Journal of Neuroscience **43**, 1074 (2023).

[3] T. F. Icard, Resource rationality, (2023).

[4] S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum, Computational rationality: A converging paradigm for intelligence in brains, minds, and machines, Science **349**, 273 (2015).

[5] C. A. Sims, Implications of rational inattention, Journal of monetary Economics **50**, 665 (2003).

[6] C. A. Sims, Rational inattention: Beyond the linear-quadratic case, American Economic Review **96**, 158 (2006).

[7] T. Berger, *Rate distortion theory: A mathematical basis for data compression* (Prentice-Hall, Inc., 1971).

[8] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).

[9] W. Schultz, P. Dayan, and P. R. Montague, A neural substrate of prediction and reward, Science **275**, 1593 (1997).

[10] H. Jeong, A. Taylor, J. R. Floeder, M. Lohmann, S. Mihalas, B. Wu, M. Zhou, D. A. Burke, and V. M. K. Namboodiri, Mesolimbic dopamine release conveys causal associations, Science **378**, eabq6740 (2022).

[11] C. R. Sims, Efficient coding explains the universal law of generalization in human perception, Science **360**, 652 (2018).

[12] N. Zaslavsky, C. Kemp, T. Regier, and N. Tishby, Efficient compression in color naming and its evolution, Proceedings of the National Academy of Sciences **115**, 7937 (2018).

[13] A. Yu, V. Ferdinand, and S. Marzen, Humans efficiently predict in a sequence learning task, in preparation (2023).

[14] A. M. Jakob and S. J. Gershman, Rate-distortion theory of neural coding and its implications for working memory, Elife **12**, e79450 (2023).

[15] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, Predictive information in a sensory population, Proceedings of the National Academy of Sciences **112**, 6908 (2015).

[16] M. Lamberti, S. Tripathi, M. J. A. M. van Putten, S. Marzen, and J. le Feber, Prediction in cultured cortical neural networks, PNAS Nexus **2** (2023).

[17] A. Hasenstaub, S. Otte, E. Callaway, and T. J. Sejnowski, Metabolic cost as a unifying principle governing neuronal biophysics, Proceedings of the National Academy of Sciences **107**, 12329 (2010).

[18] P. Mehta and D. J. Schwab, Energetic costs of cellular computation, Proceedings of the National Academy of Sciences **109**, 17978 (2012).

[19] S. Still, Information-theoretic approach to interactive

learning, Europhysics Letters **85**, 28005 (2009).
[20] L. Lai and S. J. Gershman, Human decision making balances reward maximization and policy compression, (2023).
[21] T. Malloy, C. R. Sims, T. Klinger, M. Liu, M. Riemer, and G. Tesauro, Capacity-limited decentralized actor-critic for multi-agent games, in *2021 IEEE Conference on Games (CoG)* (IEEE, 2021) pp. 1–8.
[22] W. Bialek, I. Nemenman, and N. Tishby, Predictability, complexity, and learning, Neural computation **13**, 2409 (2001).
[23] J. P. Crutchfield and S. Marzen, Signatures of infinity: Nonergodicity and resource scaling in prediction, complexity, and learning, Physical Review E **91**, 050106 (2015).
[24] S. E. Marzen and J. P. Crutchfield, Statistical signatures of structural organization: The case of long memory in renewal processes, Physics Letters A **380**, 1517 (2016).
[25] A. Celani and M. Vergassola, Bacterial strategies for chemotaxis response, Proceedings of the National Academy of Sciences **107**, 1391 (2010).
[26] S. G. Van Dijk and D. Polani, Informational drives for sensor evolution, Artificial Life 13 (2012).
[27] S. E. Marzen and J. P. Crutchfield, Predictive rate-distortion for infinite-order markov processes, Journal of Statistical Physics **163**, 1312 (2016).
[28] A. N. Burnetas and M. N. Katehakis, Optimal adaptive policies for sequential allocation problems, Advances in Applied Mathematics **17**, 122 (1996).
[29] D. B. Chklovskii and A. A. Koulakov, Maps in the brain: what can we learn from them?, Annu. Rev. Neurosci. **27**, 369 (2004).
[30] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, Planning and acting in partially observable stochastic domains, Artificial intelligence **101**, 99 (1998).
[31] F. Doshi-Velez, D. Pfau, F. Wood, and N. Roy, Bayesian nonparametric methods for partially-observable reinforcement learning, IEEE transactions on pattern analysis and machine intelligence **37**, 394 (2013).
[32] C. R. Shalizi and J. P. Crutchfield, Computational mechanics: Pattern and prediction, structure and simplicity, Journal of statistical physics **104**, 817 (2001).
[33] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, arXiv preprint physics/0004057 (2000).
[34] N. Barnett and J. P. Crutchfield, Computational mechanics of input–output processes: Structured transformations and the ϵ-transducer, Journal of Statistical Physics **161**, 404 (2015).
[35] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu, The energy–speed–accuracy trade-off in sensory adaptation, Nature physics **8**, 422 (2012).
[36] A. Uppal, V. Ferdinand, and S. Marzen, Inferring an observer's prediction strategy in sequence learning experiments, Entropy **22**, 896 (2020).
[37] N. D. Daw *et al.*, Trial-by-trial data analysis using computational models, Decision making, affect, and learning: Attention and performance XXIII **23** (2011).
[38] I. Nemenman, G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck, Neural coding of natural stimuli: information at sub-millisecond resolution, PLoS computational biology **4**, e1000025 (2008).
[39] T. Sagawa and M. Ueda, Minimal energy cost for thermodynamic information processing: measurement and information erasure, Physical review letters **102**, 250602 (2009).